Multivariate statistical assessment of the pollution sources along the stream of Kamchia River, Bulgaria

G. Mihailov*, V. Simeonov, N. Nikolov* and G. Mirinchev*****

*Department of Water Supply, Sewerage, Water & Wastewater Treatment, Faculty of Hydrotechnics, University of Architecture, Civil Engineering & Geodesy, 1 Chr. Smirnensky Blvd., 1421 Sofia, Bulgaria (E-mail: *mihailov_fhe@uecg.bg*)

**Departmen of Analytical Chemistry, Faculty of Chemistry, University of Sofia "St. Kl. Okhridski", 1 J. Bourchier Blvd., 1164 Sofia, Bulgaria

***Executive Environment Agency of Ministry of Environment & Waters, 136 Tzar Boris Blvd, 1600 Sofia, Bulgaria

Abstract Sixteen sampling sites along the stream of Kamchia River were considered as environmental objects in the multivariate statistical study aimed to identify and apportion patterns of sampling sites, latent factors responsible for the data structure and their relation to the emitter industrial and anthropogenic sources in the vicinity of the sampling sites. As variables 11 surface water parameters monitored for a long time period (up to 11 years) were used. Four main site patterns were revealed by cluster analysis (urban, rural, near-to dam and estuary) and for each site latent factors were identified and apportioned (among them "metallurgical", "food production", "winery", domestic wastes", "natural"). The relative contribution of each identified pollution source to the formation of the total concentration of each chemical species or physicochemical parameter was determined and compared to the real emitters in the region of interest.

Keywords Environmetrics; Kamchia River; water quality

Water Science & Technology **Vol 51 No 11 pp 37–43** © IWA Publishing 2005

Introduction

The contamination of the river basins is a serious environmental problem. Usually, the different approaches to reducing variability and improving information power in monitoring data intercomparison comprise normalization to a conservative component whose levels are unaffected by contaminant inputs (DeGroot *et al.*, 1976; Loring, 1990), baseline regression models based on observed covariation of elements (Hanson *et al.*, 1981; Daskalakis and O'Connor, 1995), trend studies for surface water quality to understand river water pollution impacts (Simeonov *et al.*, 2000).

However, recent studies (Simeonov *et al.*, 2003; Mihailov *et al.*, 2001; Nikolov *et al.*, 2002; Nikolov *et al.*, 2003; Stanimirova *et al.*, 1999; Simeonov *et al.*, 2002) have indicated that the application of multivariate statistical methods to estimate the monitoring data make it possible to interpret and model in a more appropriate way the complex data sets from river water monitoring.

It is the aim of the present study to assess the pollution along the stream of Kamchia River and its tributaries collecting not only monitoring data concerning water quality which was recently made and discussed in detail (Mihailov *et al.*, 2002) but also emission data from industrial and anthropogenic sources located near to the sampling sites. The assessment is carried out by the use of two major chemometrics approaches, namely cluster analysis (CA) and principal components analysis (PCA).

Methods

Sampling sites and chemical analysis

The basic pollution in the Kamchia River basin is generated by the domestic wastewaters of the settlements: Targovishte, Shoumen, Smjadovo, Veliki Preslav, Dalgopol and the industrial wastewaters of the following enterprises: "Vinex-Preslav"AD, town of Veliki Preslav, ET "Zlatno runo", town of Veliki Preslav, "Vinarska izba-Han Krum" EAD Han Krum village, Veliki Preslav community, "Energia"AD, town of Targovishte, "Bjal potok"OOD, dairy of Davidovo village, Targovishte community, "Bramas-96"AD(incinerator), town of Shoumen, PHJ"Brothers commers" AD, town of Shoumen, "Hybrid center of pig-breeding", town of Shoumen, "Alkomet", town of Shoumen (EEA of MoEW, 2003). The water quantities (Q, Mm^3/year) corresponding to the different branches are as follows:

Domestic wastewaters: Shumen – 9.30, Targoviste – 4.42, Smiadovo – 0.35, Veliki Preslav – 0.79, Dalgopol – 0.41. Total: Q = 15.27 (Sew. Design Standards, 1994).

Ferrous/non-ferrous metallurgy:

- "Alkomet", town of Shoumen — Q = 0.206;

Food, wine and tobacco industries:

- "Vinex-Preslav" AD, town of Veliki Preslav — Q = 0.077;
- ET"Zlatno runo", town of Veliki Preslav — Q = 0.0005;
- "Vinarska izba-Han Krum" EAD Han Krum village, Veliki Preslav community — Q = 0.0303;
- "Bjal potok" OOD, dairy of Davidovo village, Targovishte community — Q = 0.008
- PHJ"Brothers commers" AD, town of Shoumen — Q = 0.5;
- Total for the branch: Q = 0.62.

Other industries:

- "Energia" AD(battery production), town of Targovishte — Q = 0.928;
- "Bramas-96"AD(incinerator), town of Shoumen — Q = 0.027;
- "Hybrid center of pig-breeding", town of Shoumen — Q = 0.072.
- Total for other industries: Q = 1. 027.

The total industrial wastewater quantity entering Kamchia River during 2002 is:

Q = 1.85.

The sampling sites used are indicated by numbers as shown in Figure 1. These sites include urban and rural environments, tributary sites, near-to-dam collectors, and river estuary sites. The sampling period varied between 5 and 10 years for the different sites and included such important parameters of the surface water quality as pH, dissolved oxygen, biological oxygen demand (BOD_5), dissolved matter, suspended (non-dissolved) solids, chloride, ammonium – nitrogen, nitrate-nitrogen, nitrite, phosphate and iron.

The chemical analysis performed includes standard analytical methods as routinely applied in the control laboratories of the monitoring net (potentiometry, titrimetry, gravimetry, and spectrophotometry). The sample preparation and sample measurements are described in detail elsewhere (Bulgarian standards for water quality, Sofia, 1985).

Multivariate statistical methods

Cluster analysis. The notion cluster analysis encompasses a group of methods, which are mainly applied for finding and visualizing structures and similarities within sets of data (Massart and Kaufman, 1983; Einax *et al.*, 1997). The most common similarity measure or similarity distance between objects described by variables is the Euclidean distance. One important aspect of performing cluster analysis is to ensure the comparability of

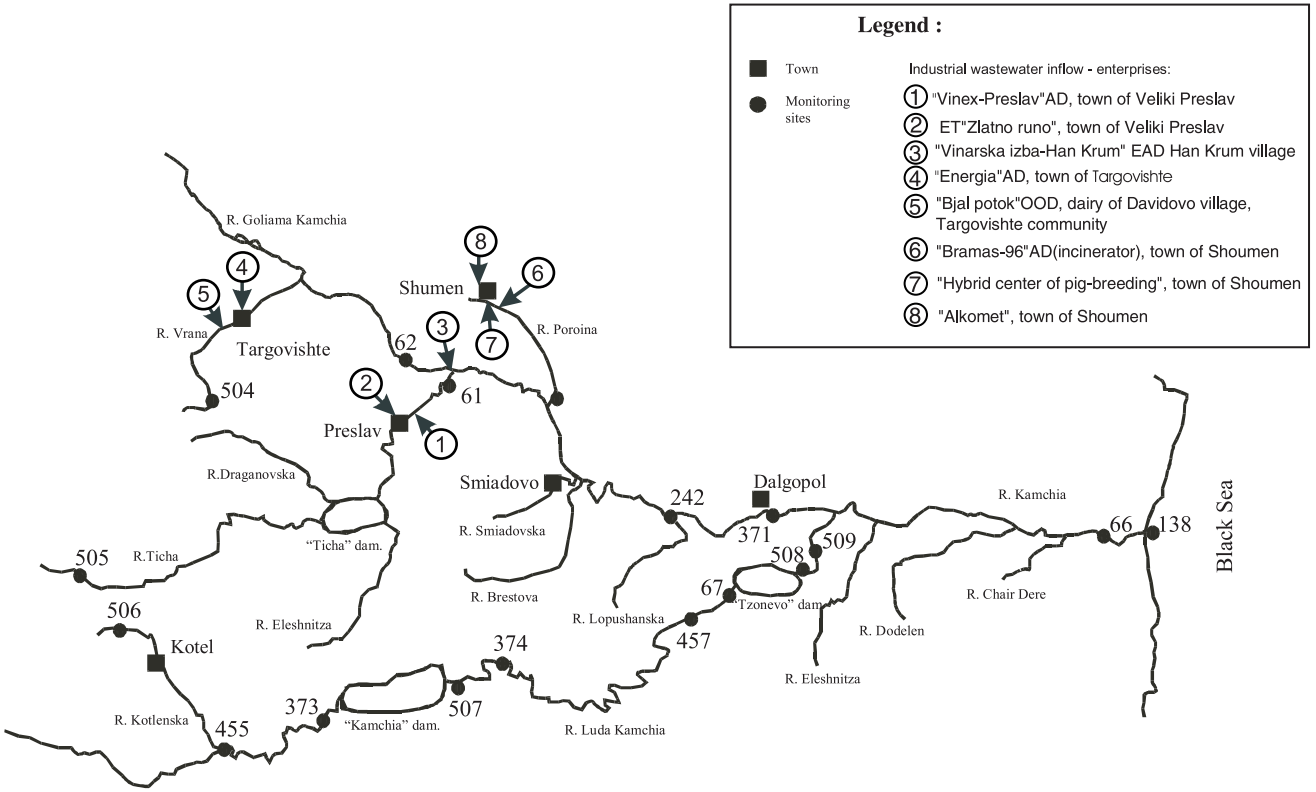

Figure 1 Location of sampling sites along the stream of Kamchia River

the variables and that is why a data preprocessing by autoscaling or z-transformation is necessary.

After selecting a measure one has to choose a clustering algorithm (way of linkage of the objects). Hierarchical and non-hierarchical techniques are known applying the methods of single linkage, average linkage, method of Ward etc. All objects are then linked with each other in some hierarchy depending on the distance between them. The similar objects tend to form groups of similarity (clusters), which have to be tested for significance (Sneath, 1980) and then an appropriate interpretation and graphical output is possible.

Principal components analysis (PCA). PCA is aimed at finding and interpreting hidden complex, and possibly, casually determined, relationship between variables in the data set. Correlating variables are converted to so-called factors, which are themselves non-correlated. The central task in PCA is to reduce the original data matrix X (m objects and n variables, or $m \times n$) to the following component parts of factor loadings A and factor scores F matrices

$$X = A.F,$$

as the number of factors is theoretically equal to the number of variables.

A linear combination of different factors in the matrix A with factor scores in the matrix F can reproduce the data matrix X. These factors are new synthetic variables and represent a certain quality of the variables from the data set. They explain the total variance of all variables in a descending order and are themselves non-correlated.

Results and discussion

The data set available includes 16 sampling sites (objects) and 900 analytical results (EEA of MoEW, 2003). Average annual concentrations as variables are used. The distribution of the results among the sites is not uniform due to the various terms of monitoring: in some cases it comprises 10 years, in others – only 5. In Table 1 a short characteristic of each site is given along with the number of samples and years of monitoring.

It was of substantial interest to estimate the level of similarity between the sampling sites. Any reliable proof for similarity between sites of similar location (e.g. urban or rural, or tributary etc.) could confirm a hypothesis for the presence of "site patterns" along the stream. Indeed, several clusters are obtained in the hierarchical dendrogram

Table 1 General description of the sampling sites

Site	Monitoring	Samples	Type	Significant emitting sources
61	1992–2001	109	urban	domestic wastes, winery, food production
62	1992–2001	113	urban	domestic wastes, winery, food production, metallurgy
63	1992–2001	172	urban	domestic wastes, winery, food production, metallurgy
66	1992–2001	100	rural	domestic wastes
242	1992–2001	86	rural	domestic wastes, farming
371	1993–2001	84	urban	domestic wastes, farming
373	1993–2001	83	urban	domestic wastes, farming
374	1993–2001	40	rural	domestic wastes, farming
457	1994–2001	63	rural	domestic wastes, farming
504	1998–2001	11	urban	domestic wastes, power station, food production
505	1998–2001	11	rural	farming
506	1998–2001	10	rural	farming
507	1998–2001	17	dam	natural sources
508	1998–2001	15	dam	natural sources
509	1998–2001	8	dam	natural sources
138	1992–2001	78	estuary	natural sources

(Figure 2) which is constructed by the Ward method of linkage with squared Euclidean distance as similarity measure. For the cluster analysis average annual concentrations as variables are used.

Thus, each site is represented by several objects, which depend on the number of monitoring years, e.g. 10 objects for each of the sites 61, 62, 63, 66, 242, and 138; 9 objects for sites 371, 373 and 374; 8 objects for site 457 and 4 objects for each of the sites 504, 505, 506, 507, 508, and 509. In cluster 1 dominant objects from sites 61, 62, 63, 371, 373, and 504 are found, which reflects the idea of an "urban site" pattern.

All of these sites are located near to urban regions and show enhanced concentrations of ammonium and nitrate ions. In cluster 2 are included dominant "rural pattern" sites (66, 242, 374, 457, 505, 506) with higher values of pH, lower values of BOD_5 and iron. Cluster 3 involves dominant sites located in the vicinity of dams (507, 508, 509) but as outlet of the dams.

It may be assumed that the water collection in the dams plays a purification role and these sites are very similar to the rural ones but still different ("dam outlet" pattern). Finally, site 138, which is a typical estuary site, is located differently as an outlier ("estuary" site).

Next important step of the multivariate statistical analysis was to compare the patterns obtained with the information about emitting sources in the region. As mentioned in the introductory part, the main polluting sources along the steam of Kamchia River are:

- Metallurgical industry: high correlation between iron, non-dissolved matter (suspended matter), ammonium and phosphate;
- Food production: high correlation between nitrate and dissolved matter and BOD_5;
- Wineries: high correlation between non-dissolved matter, dissolved oxygen, BOD_5;
- Organic wastes: high correlations between ammonium, nitrates, sulfates, nitrites, permanganate oxidation;
- Inorganic wastes: high correlations between chlorides, pH, dissolved matter;
- Farming: high correlation between nitrates, phosphates and ammonium;
- Natural sources: high correlations between dissolved oxygen and dissolved matter.

Keeping in mind this idea of polluting sources, PCA of each sampling site was performed in order to identify emitting sources and compare them to really existing industries and anthropogenic waste inlets along the stream. In Table 2 the results of the PCA carried out are presented along with the location of real emitters in the vicinity of each sampling

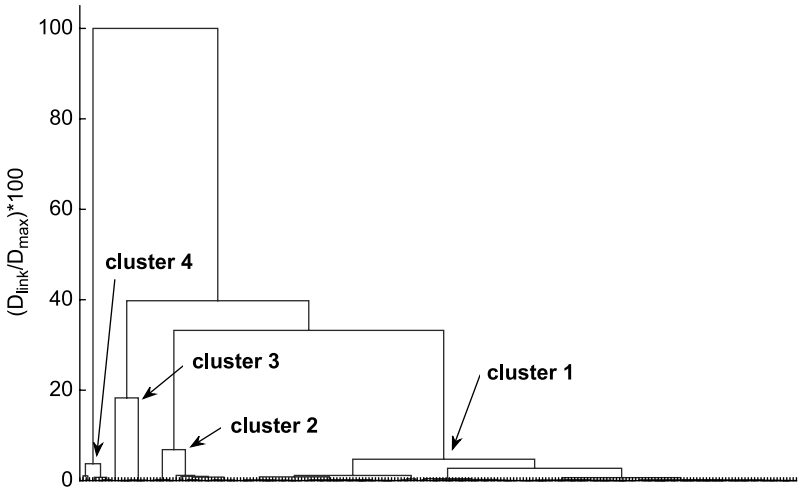

Figure 2 Hierarchical dendrogram of the sampling sites (Ward method of linkage)

Table 2 Pollution source identification for each site by PCA

Site	PC1	PC2	PC3	PC4
61	BOD_5, NO_3^- DM Food production 34.2%	DO, NO_2^-, Fe Winery 28.3%	Cl^-, PO_4^{3-}, pH Inorg. wastes 16.1%	NH_4^+, SS Org.wastes 10.4%
62	Fe, BOD_5, SS, NO_3^- Metallurgy 42.4%	DO, NO_2^- Winery 18.9%	Cl^-, PO_4^{3-}, pH Inorg. Wastes 10.4%	NH_4^+, DM Org.wastes 9.1%
63	Fe, BOD_5, SS, NO_3^- Metallurgy 37.2%	DO, NO_2^- Winery 21.9%	Cl^-, PO_4^{3-}, pH Inorg. Wastes 10.4%	NH_4^+, DM Org.wastes 6.1%
66	DM, DO, BOD_5, pH Natural factor 47.2%	Cl^-, PO_4^{3-} NO_3^- Inorg. Wastes 16.1%	NH_4^+, SS, NO_2^- Org.wastes 9.7%	–
242	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 36.1%	Cl^-, PO_4^{3-}, pH, Fe Inorg. Wastes 22.1%	DO, BOD_5, DM Org. wastes 12.4%	–
371	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 41.1%	Cl^-, PO_4^{3-}, pH, Fe Inorg. Wastes 18.1%	DO, BOD_5, DM Org. wastes 15.4%	–
373	NH_4^+, NO_3^-, NO_2 Farming factor 36.6%	Cl^-, PO_4^{3-}, pH, Inorg. Wastes 22.1%	DO, BOD_5, DM Org. wastes 14.4%	SS, Fe Natural factor 8.1%
374	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 33.1%	Cl^-, PO_4^{3-}, pH, Fe Inorg. Wastes 28.7%	DO, BOD_5, DM Org. wastes 13.9%	–
457	Cl^-, PO_4^{3-}, pH, Fe Inorg. Wastes 28.1%	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 26.1%	DO, BOD_5, DM Org. wastes 19.4%	–
504	BOD_5, NO_3^- DM Food production 32.6%	SS,NH_4^+, NO_3^-, PO_4^{3-} Power station 27.9%	DO, Cl^-, pH, Fe Domestic wastes 19.2%	–
505	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 42.1%	DM, DO, BOD_5, pH Natural factor 27.2%	Fe, Cl^-, PO_4^{3-} Domestic wastes 13.3%	–
506	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 34.9%	DM, DO, BOD_5, pH Natural factor 21.3%	Fe, Cl^-, PO_4^{3-} Domestic wastes 17.2%	–
507	DM, DO, BOD_5, pH Natural factor 51.2%	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 34.9%	–	–
508	DM, DO, BOD_5, pH Natural factor 44.3%	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 33.3%	–	–
509	DM, DO, BOD_5, pH Natural factor 48.1%	NH_4^+, NO_3^-, NO_2^-, SS Farming factor 31.6%	–	–
138	DM, DO, BOD_5, pH NH_4^+, NO_3^-, NO_2^-, SS "Estuary factor" 60.2%	–	–	–

Note: DO means dissolved oxygen; SS – susspended solids; DM – dissolved matter

site. The source identification by PCA is performed using the Varimax rotation mode of PCA. As statistically significant, loadings higher than 0.70 are taken into account (Malinowski, 1991).

Each latent factor (PC) is presented in the table with the correlated variables (statistically significant factor loadings), with its conditional name and percentage of explained variance.

It might be seen that each sampling region offers its specific pattern of pollution. The identification of the polluting sources by PCA reveals that in some cases the correlation between the water parameters varies from site to site indicating in one case a certain latent factor, in another case – quite different. This is no surprise as the water quality at the different sites also differs due to the variability of the emitters. A very important and specific validation of the PCA models for any sampling site is not only the percentage of the explained total variance (over 75% at almost all sites) but the very exact coincidence between the identified latent factors and the real emitters in the vicinity of the sites. Thus, the approach applied offers a simple opportunity for source identification and apportioning.

Conclusion

The study carried out makes it possible to gain specific information about polluting sources along the stream of Kamchia River. The known industrial wastes (from metallurgy, food production, wineries, domestic inorganic and organic inlets to the river flow, etc.) as well as natural sources are additionally recognized and apportioned by the use of principal components analysis. Different patterns of sites are proved by the cluster analysis, which adds information to the whole scheme of monitoring. The multivariate statistical analysis enables better decision-making and problem solving in a delicate environment.

References

Bulgarian State Standards for Water Quality (1985), EN and ISO.

Daskalakis, K. and O'Connor, T. (1995). *Mar. Environ. Res.*, **40**, 389–398.

DeGroot, A., Salomons, W. and Allersma, E. (1976). Processes affecting heavy metals in estuarine sediments. In *Estuarine Chemistry*, Burton, J. and Liss, P. (eds), Academic Press, New York.

Einax, J.W., Zwanziger, H. and Geiss, S. (1997). *Chemometrics in Environmental Analysis*, VCH Weinheim.

EEA of MoEW (2003), Information-control system for identification of the wastewater status according to Regulation No. 5, Water Low, Work Shop, Sofia

Hanson, P., Evans, D. and Colby, D. (1981). *Mar. Environ. Res.*, **36**, 237–266.

Loring, D. (1990). *Mar. Chem.*, **29**, 155–168.

Massart, D.L. and Kaufman, L. (1983). *Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, Wiley, New York.

Malinowski, E. (1991). *Factor Analysis in Chemistry*, Wiley, New York.

Mihailov, G., Simeonov, V., Nikolov, N. and Mirinchev, G. (2001). *Toxicol. Envir. Chem.*, **83**, 1–12.

Mihailov, G., Simeonov, V., Nikolov, N. and Mirinchev, G. (2002). *Water Sci. & Technol.*, **46**(8), 45–52.

Nikolov, N., Mihailov, G., Simeonov, V. and Mirinchev, G. (2002). *Chem.Eng. Ecol.*, **9**, 1532–1539.

Nikolov, N., Mihailov, G., Simeonov, V. and Mirinchev, G. (2003). *Chem. Eng. Ecol.*, **10**, 763–771.

Sewerage Design Standards (1994), Sofia, MRDPW.

Simeonov, V., Einax, J.W., Stanimirova, I. and Kraft, J. (2002). *Anal. Bioanal. Chem.*, **374**, 898–905.

Simeonov, V., Stefanov, S. and Tsakovski, S. (2000). *Microchim. Acta*, **134**, 15–21.

Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. and Kouimtzis, Th. (2003). *Water Res.*, **37**, 4119–4124.

Sneath, P.H.A. (1980). *Data Analysis and Informatics*, North-Holland, Amsterdam.

Stanimirova, I., Tsakovski, S. and Simeonov, V. (1999). *Fresenius J. Anal. Chem.*, **365**, 489–493.